Comparison of clustering of 16S and 5S RNA genes of bacteria by their triplet composition

Iuliia Ovchinnikova*, and *Vyacheslav* Leonov

Siberian Federal University, 79, Svobodny av., Krasnoyarsk, 660041, Russian Federation

Abstract. The study of biological macromolecules such as RNA and the search for novel methods to analyze them is a crucial task due to their fundamental importance in all living organisms. Currently, analyzing the information stored in the genome is a complex process, and it is essential to find new ways to comprehend the functions and structures of genes and their interactions. This paper analyzes the clustering of bacteria based on the triplet composition of 5S and 16S RNA genes using elastic maps derived from frequency dictionaries of triplets. We have created indexed databases of 16S and 5S RNA genes and performed a comparative analysis of their clustering. We have determined the taxonomic composition of the identified clusters and analyzed the relationship between the structure of these clusters and the taxonomy of the bacteria.

1 Introduction

The study of the relationship between the structure of biological macromolecules, specifically bacterial ribosomal RNA (rRNA), and the classification or taxonomy of their host organisms, is an important field of research for molecular biologists and bioinformaticians. While the classic focus has been on 16S RNA sequences [1-2], it is also crucial to explore the connection between structure and classification for other types of RNA.

In this study, we examined the correlation between RNA structure and taxonomy by analyzing bacterial 16S and 5S RNA genes.

The development of technology has made it possible to study RNA using various methods. One such method is to determine the structure of the genome using frequency dictionaries of triplet sequences [3]. This method has already been applied to bacteria in general [4]. The use of frequency dictionaries allows us to identify spatial structures that provide information about gene function, taxonomic classification, and the relationship between structure and function or taxonomy.

The main goal of this research is to cluster multidimensional data using the elastic map [5], a method for visualizing and solving problems. This work aims to identify similarities and differences in bacterial clustering based on triplet composition in 5S and 16S RNA genes.

* Corresponding author: july.l4o6@mail.ru

2 Materials and methods

The SILVA database was chosen for the study of 16S rRNA and 5S rRNA genes, as it is well-suited for fast and accurate taxonomic classification of rRNA sequences. To study clustering, we downloaded files with data on 16S and 5S rRNA gene sequences from the open SILVA database. However, all databases, including SILVA, have a problem in that the number of organisms in lower taxa varies greatly, which can cause distortion in clustering results. To address this issue, we indexed the databases by aligning the number of represented taxa. We have proven that different indexing options for both 16S rRNA [2] and 5S rRNA [6] do not significantly affect the pattern of distribution. The table 1 shows the indexed database.

Table 1. Composition of the indexed database of 5S and 16S RNA genes.

Phylum	**16S**	**5S**
Actinobacteriota	24	20
Bacteroidota	695	70
Firmicutes	1201	1890
Verrucomicrobiota	189	172

The distribution of genes was studied by transforming them into a mathematical object called a frequency dictionary of triplets.

Since the frequencies of all triplets are linearly dependent, one triplet should be excluded from the analysis. It was decided to exclude the triplet with the minimum value of the standard deviation from the analysis, since it makes the least contribution to the distinctiveness of genes. In this study, the CAC triplet was excluded.

To perform gene clustering it was chosen to use the elastic map method because it offers advantages in data interpretation, visualization and dimensionality reduction.

The elastic map method helps to identify heterogeneities in the distribution of data for identify clusters. In this case, identifying clusters on an elastic map means that there is some structure in the data. Clusters are identified based on local density. To determine it on an elastic map, each point is supplied with a dome-shaped function with a maximum at this point, for example, a Gaussian function (1).

$$f_j(r) = A \cdot exp\left\{-\frac{(r-r_j)^2}{\mu^2}\right\} \quad (1)$$

where rj – the coordinate of the jth point;

A – a multiplier that is the same for all points;

μ – the half–width of this function, which is completely analogous to the standard deviation for the case of a normal random variable distribution;

Then we determine the local density by summing the functions over all points (2).

$$F(r) = \sum_{j=1}^{N} exp\left\{-\frac{(r-r_j)^2}{\mu^2}\right\} \quad (2)$$

where N – the number of points.

To visualize the obtained data was used VidaExpert software.

3 Results and discussion

Elastic maps were made by frequency dictionaries of triplets in VidaExpert software, determined by local density on an elastic map in internal coordinates (Fig. 1).

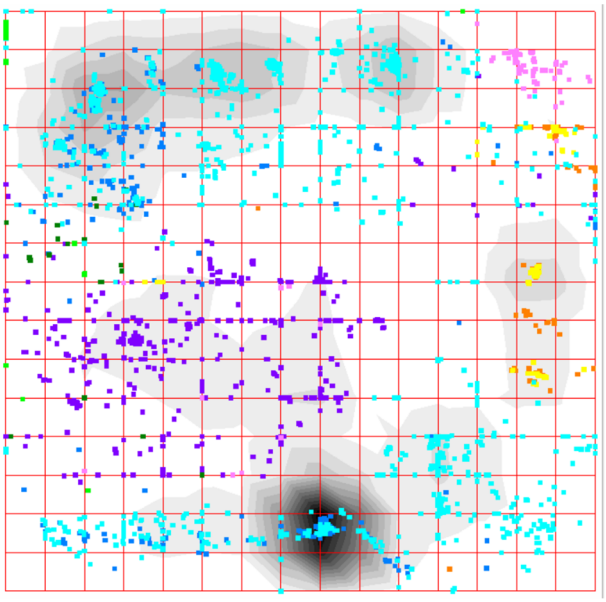

Fig. 1. Distribution of bacterial 16S and 5S RNA genes on an elastic map taking into account taxonomy.

The color gradient reflects the local density: the denser the distribution of points, the darker the color of such a section of the map. A correlation radius of $\mu = 0.25$ was used to calculate the local density.

The points on the elastic maps indicate the frequency dictionaries under study (Table 2). These frequency dictionaries were studied in relation to the taxonomic affiliation of each gene to a specific organism (bacterial species).

Table 2. The color coding of the phylum of bacteria.

Phylum	16S rRNA	5S rRNA
Actinobacteriota		
Bacteroidota		
Firmicutes		
Verrucomicrobiota		

It is clearly seen that the RNA genes of different phylum are located on the elastic map (Fig. 1) in a completely non-random manner which should be understood as a strong preference in the distribution of genes by clusters.

When studying the joint distribution of bacterial 16S and 5S rRNA genes, we identified various clusters. These clusters are determined by the local density of points in the frequency space of triplets.

It can be selected a different number of clusters for the analysis. We have highlighted the areas with the highest local density (visually represented by darker areas on the map) as clusters.

In this manner, there were identified 6 clusters based on local density (Fig. 2).

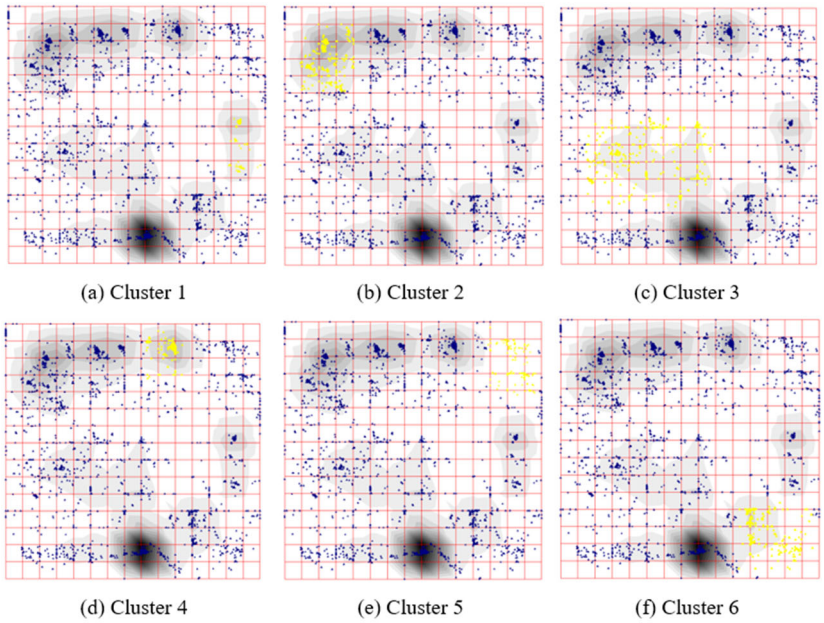

Fig. 2. Distribution of 5S and 16S RNA genes into 6 clusters (yellow points).

Using the capabilities of the ViDaExpert software, there was made a diagram that shows the percentage of phylum and genes in clusters (Fig. 3).

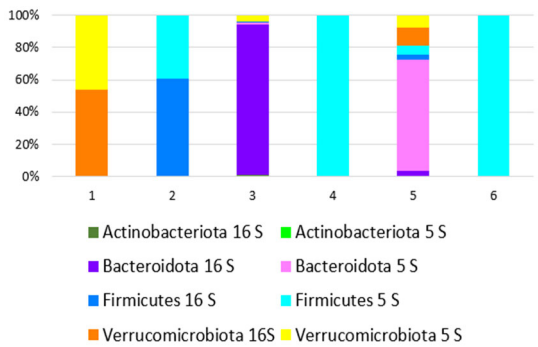

Fig. 3. Percentage of the phylum and genes in clusters.

The expected result of the gene clustering study was that the identified clusters would include mostly taxonomically related organisms that would contain both 16S and 5S RNA genes. Clusters 1 and 2 provide direct evidence for this assumption (Fig. 3).

However, we can conclude that the Bacteroidota and Firmicutes are special groups, since the representatives of these phylums are formed by several clusters.

It is possible to identify several reasons for this behavior of genes. The first reason can be attributed to the biological properties of microorganisms in these environments. Additionally, such behavior on the map can be justified by errors in the database or mistakes made by researchers in identifying species.

The first reason suggests that the researchers who compiled the database made some errors. The second possibility indicates errors in the identification and classification of certain microorganisms.

Moreover, the effect of splitting genes of one phylum into several clusters may be due to the type of elastic map chosen for analysis. It is quite possible that using a sphere will result in the merging of two clusters located in different parts of the square into one.

4 Conclusion

Within the framework of this study, we investigated the relationship between the structures of the clustering of 5S and 16S RNA genes, and compared the clustering of these genes by triplet composition in bacteria. To do this, we solved the following tasks:

1) an indexed database of 16S and 5S RNA gene sequences and a family of frequency dictionaries for each sequence included in the database have been created.

2) a comparative analysis of the clustering patterns of 16S and 5S RNA was performed, which showed that these genes are not randomly distributed among bacterial species. Instead, there is a strong preference for grouping genes into clusters.

3) the taxonomic composition of the identified clusters has been determined. Special attention should be paid to those departments whose genes have diverged into different clusters. The reasons for this discrepancy may be due to several factors, including errors in filling out the database, mistakes made by researchers in identifying species, actual biological features of these taxa, or the type of elastic map chosen for analysis.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

References

1. M. Kim, J. Chun, Methods in Microbiology, **41**, 61-74 (2014)
2. Y. W. Tang, C. W Stratton, X. Y. Han, *Advanced techniques in diagnostic microbiology* (Springer, 2006).
3. A. Teterleva, V. Abramov, A. Morgun, I. Larionova, M. Sadovsky, LNBI **13346**, 205-215 (2022)
4. A. Gorban, T. Popova, A. Zinovyev, Physica A: Statistical Mechanics and its Applications, **353**, 365-387 (2005)
5. A. N. Gorban, A. Yu. Zinovyev, International Journal of Neural Systems, **20**, 219-232 (2010)
6. I. Ovchinnikova, M. Sadovsky, V. Sokolov, ITM Web of Conferences **59,** 04014 (2024)